NITED NATIONS
CONOMIC
ND
OCIAL COUNCIL

Distr.
LIMITED

ST/ECLA/CONF.32/L.9
2 May 1968

ORIGINAL: ENGLISH

SEMINAR ON THE ORGANIZATION AND
CONDUCT OF POPULATION AND HOUSING
CENSUSES FOR LATIN AMERICA

Santiago, Chile, 20-31 May 1968

Organized by the United Nations Economic
Commission for Latin America, Statistical
Office and Office of Technical Co-operation,
with the collaboration of the Inter-American
Statistical Institute, the Latin American
Demographic Centre and the United States
Bureau of the Census

ELECTRONIC PROCESSING OF CENSUS DATA

PART I

Machine Processing of Census Data*

Contents

* Prepared by the Secretariat of the Economic Commission for Asia and the Far East for the Seminar on The Organization and Conduct of Population and Housing Censuses held at Bangkok, Thailand, 24 November-1 December 1967, for the countries of the ECAFE region, and issued at ECAFE as United Nations document E/CN.11/ASTAT/SPHC/L.8, 2 November 1967.

/...

I. INTRODUCTION

1. With few exceptions the results of the 1960 round of censuses of population in the ECAFE region were tabulated with conventional punched-card machines. Judging by the rate at which the statistical work in the region has become computer-oriented in recent years, it is evident that all but a small number of countries are planning to use computers for processing the 1970 round of censuses. For those few statistical offices which will not have access to a computer in their own countries, it may be possible to consider the use of computer facilities of a nearby country or of the proposed ECAFE regional computer centre. For organization and planning of the forthcoming censuses it is, therefore, assumed in this paper that electronic data processing techniques will be used almost exclusively throughout the region.

2. Considerable documentation giving guidance in the use of punched-card methods in processing census data was prepared for the 1960 population censuses. In particular, many of the principles enunciated in the informative publication "Handbook on Data Processing Methods" published jointly by the United Nations and the United Nations Food and Agriculture Organization in 1959 are directly applicable to electronic data processing. Some materials designed to aid in computer processing of censuses have also been prepared by the countries.

3. Among the series of activities planned for the 1970 round of population and housing censuses it has been proposed to hold seminars dealing specifically with the technical considerations in the application of computers to census data processing. For the ECAFE region, this is being considered for the first half of 1969. The documentation from this and similar meetings, and the services of advisers from international and other organizations with previous experience in computer processing of population census data will be available for assistance to ECAFE countries in preparing for the 1970 round. The discussion in the present paper will be confined, therefore, to general principles and aspects of planning for computerized processing of the population and housing censuses.

/4. For

4. For many countries of the region the 1970 censuses will represent an initial exposure to the use of computer techniques. It may, therefore, be in order to give a brief description of the main characteristics of computer processing, particularly as applied to census work.

II. ADVANTAGES OF COMPUTER PROCESSING

5. The use of punched-cards for recording statistical data and sorting, counting and tabulating machines for compiling and summarizing the information recorded in the cards was developed originally in response to the needs for tools to handle the massive task of processing the census. The advent of the electronic computer has provided the potential for far greater speed, versatility and accuracy than the previous methods of processing census work. At the present time, electronic equipment is used extensively for mass processing of statistical programmes in virtually all modern governmental statistical agencies throughout the world.

6. In general, the principal attributes of electronic data processing which offer impressive superiority over conventional punched-card methods include:

(i) Tremendously improved speed in computations and transfers of data.

(ii) Memory storage which provides the capacity for "remembering" and retrieving at high access speeds the data stored in the memory.

(iii) The capacity for automatic execution of sequential instructions.

(iv) The capability of making logical decisions automatically during the course of processing.

(v) Considerably improved accuracy and self-checking features.

The combination of these characteristics of electronic computing equipment thus renders it possible to solve complex data-handling problems in one continuous process at electronic speeds, as opposed to the segmented handling of data as performed in the conventional punched-card system. The capacity to process vast quantities of data and the flexibility of the new equipment represent a qualitative difference from the earlier devices that make practicable new techniques in statistical operations.

/7. For

7. For example, the computer is commonly used to excellent advantage in census work for editing and correction of the data prior to or in combination with tabulation. The computer can be directed to automatically assign, or impute, values when information on personal or housing characteristics is omitted in the questionnaire or when certain information reported (or encoded) is inconsistent or impossible. Most computers are capab capable of applying to the data all the rules that a group of clerks would use in detecting and reconciling inconsistencies and inferring missing information. A common situation encountered, for example, is when a given characteristic is inconsistent with other characteristics reported for that person. The general procedure under these circumstances may be to draw a value of the characteristic from a distribution of that characteristic for the appropriate sub-group in the population. Thus, a person who is reported as a male relative of the household head, but for whom marital status is not reported, may be assigned by the computer a marital status from a distribution of marital status for male household head relatives in the same age group.

8. Similar procedures may be used for the detection of missing or impossible data. In many ways, the computer performs these editing tasks strikingly better and faster than when they are done by manual inspection. It will apply the editing rules individually to each person in the enumeration. It will do this with strict consistency and uniformity, and can be directed to keep a record of how often each such rule was applied, thus providing a measure of the quality of the enumerated data.

9. The use of a computer in census processing also provides the potential for development of more detailed or more informative statistics. The flexibility of the computer makes possible far greater utilization of data for derived statistics and by-products that become available for the first time. Key ratios and relationships that formerly could not be obtained without costly reprocessing of data or vast amounts of manual computation can be programmed as an incidental by-product of computer tabulations. Data can more easily be rearranged by more variables of classification

/and more

and more detailed cross-classifications which, by other methods, would be prohibitive in time and cost. It is also technically possible to combine and relate census data with data from other sources concerning the same area of inquiry. Thus, it may be possible to match and combine census data with those from other sources which, for one reason or another, cannot be collected in the census. For example, vital registration data might be used for inter-censal updating of some of the population characteristics compiled in the current census.

10. Another advantage of processing census data on computers is that it often permits preparation of the final tables in suitable form directly from the machine-prepared tables for printing by offset printing processes. With earlier mechanical tabulating equipment, the results usually cannot be prepared in a form in which they can be sent to a printer. It is generally necessary to do a substantial amount of hand posting from the machine sheets and then to type the posted tables in final publication form. Preparation of the final tables by computer for direct printing eliminates the need for manuscript, type-setting, and proofreading. On this basis, the final results of the census could be released to the public many months earlier than with conventional printing methods.

11. Where physical space and storage facilities are at a premium, as is commonly the case throughout the region, computers require considerably less floor space than that required for conventional equipment. In countries using tape computer systems, the storage of census records on magnetic tape requires only a fraction of the space normally needed for card storage.

12. There are, of course, some disadvantages and difficulties involved in the transition to computers. The use of this kind of equipment brings new problems which require that the processing of census data be organized in a more systematic manner than for conventional punched-card equipment. One requirement in applying computers is that very detailed and explicit instructions must be provided to the computer before it will proceed with any operation. For example, in the machine-edit procedures, the machine must be instructed to provide for every relevant combination of entries it may

/encounter.

encounter. Failure to consider all contingencies may result either in bringing the computer to a halt every few minutes or in rejecting large numbers of records because of some flaw in the data for which adequate provision was not made in the computer programme. For this reason, the programming and supervisory staff must be highly competent and well trained. Personnel with these attributes are difficult to find and much in demand. To attract qualified personnel it is often necessary to offer inducements in the form of premium salaries and special allowances. Preparation of physical facilities for the computer are more demanding than with conventional punched-card equipment, requiring air-conditioning, dust control, special flooring, etc.

13. Above all, it must be recognized that the availability of sophisticated data processing equipment will not, in itself, insure efficient and accurate census results. Successful computer application demands a much higher degree of precision and discipline in planning the census processing operations than was required with previous techniques.

III. NEED FOR EARLY PLANNING

14. The methods and equipment used in processing are directly related to the objectives and results of the census, and have a broad impact in virtually all operations. Each step in the census is closely linked with the previous operation and, in turn, influences subsequent operations. Thus, it is of paramount importance that the data processing function be represented in the planning and organization of the census from the very beginning as an integral member of the over-all census planning group.

15. The data processing representative in the planning group should, for example, actively participate in the design of the questionnaire to insure that the final design meets the requirement for efficient processing. He should assist in the development of control procedures, editing and coding specifications, the tabulation requirements, schedules, and other factors so that they may be considered with realism by the census planning authorities.

16. The introduction of the computer to census processing, by its nature, increases the time necessary for preparation for statistical processing. A great premium is therefore placed on careful advance planning. In the past, those responsible for processing the data, have, in many cases, been forced to do their detailed planning by improvisation and on the spur of the moment. This type of planning when using electronic computers in not only more difficult but perilous from the point of view of the end results. Computer programming is extremely intricate and requires planning of a high order and painstaking attention to detail. The cost, effort and time involved in changing computer procedures makes it undesirable to change those once established. Thus, if the desired time schedules are to be achieved, it requires that complete and final tabulation plans and specifications must be available to technical programmers at a much earlier date than was required by previous processing systems. The possibilities for flexibility and improvisation of tabulation plans under conditions of tight schedules are extremely limited. This, of course, suggests that over-all planning of the census should be accomplished well in advance. Many countries, even those with long histories of census taking, devote five or more years to active census planning. The United States, with 180 years of census experience, has been actively planning the 1970 census since before 1954.

IV. INFLUENCE OF EQUIPMENT ON PLANNING

17. The capacity, speed, and configuration of the computer to be used for processing the census is naturally the factor having the greatest bearing on the processing plan. Equipment may range from low speed card-operated computers with very limited core storage to installations with large storage capacities and a wide range of peripheral equipment. These considerations will influence and, in many cases, dictate the entire approach to the census plan, including design of the enumeration form, field work, coding, input preparation, editing and tabulation. The impact of equipment considerations on the whole plan for the census may best be illustrated by the far-reaching effects which the use of optical scanning techniques in the enumeration would have on all census operations.

18. For planning purposes, therefore, it is essential to know, at least in general terms, the equipment on which the census will be processed. The specific distinctions that are needed for planning are as between equipment systems operating with (i) card only; (ii) card and magnetic tape; and (iii) card, tape and optical reading devices.

(i) Card-only

19. Use of computers with only card input and output capabilities will speed up the statistical compilation work as compared with the use of conventional punched-card machines; but it will actually represent a minimum change from the earlier methods. Computeres provide greater computational speed and data-edit capabilities than conventional punched-card machines, but the magnitude of the key punch work in preparing the input cards remains the same as previously. However, the higher speed card reading devices (as much as 800 cards per minute versus approximately 150 per minute with conventional tabulators) permit each pass on the computer to be made more quickly while the memory storage permits tabulation of more variables on each pass. The memory storage facility of these machines also provides the capability for effectively editing the recorded data for invalid codes, inconsistency and reasonableness. However, sorting, collating and other operations on the cards much continue to be done on punched-card equipment in the conventional manner.

(ii) Card and magnetic tape

20. The most common type of equipment on which the forthcoming round of censuses will be processed within the region is likely to be a medium scale system with card and magnetic tape devices, printer, and possibly magnetic disc or drum storage peripherals. After the cards have been punched, converted to tape, edited and corrected where necessary, it is possible on such installations to sort and tabulate the results with considerable speed. In these cases, the cost and delay of manually punching the data before it can be transferred to magnetic tape for high speed processing in the computer must still be considered in the processing plan. The impact of such computers will be in the

/improved

improved speed of computation and final tabulation of the data. The tabulation work, however, may not be begun until a "clean" tape is achieved and, therefore, special emphasis in census planning should be placed on expediting and controlling the quality of the card punching, card-to-tape, and editing operations.

(iii) Card, tape and optical reading devices

21. Optical scanning systems are automatic methods of preparing punched cards or magnetic tape records from the enumeration form and are of relatively recent development. Some countries in the ECAFE region are considering the use of these devices in the 1970 censuses. In these cases, the planning for allocation of resources will be affected by elimination of the need for large scale punching operations. On the other hand, additional resources and effort will be required to assure the quality of field work necessary to obtain a high percentage of machine-readable enumeration forms. Optical scanning may be by mark-sensing or character recognition equipment. In the latter category, although developments have been encouraging, the techniques has not yet been advanced to the point at which it may be generally considered for population censuses. Character-reading equipment which will reliably read the hand-written or non-standardized characters encountered in this work is not yet suitable for general purpose application, but equipment for reading position marking has been and can be successfully applied. Further experience with these techniques, in areas where the field work is more amenable to control, is desirable.

V. SYSTEM DESIGN AND PROGRAMMING

22. Considerable system design and computer programming work will be required for the machine-edit and tabulation phases of the census. There is a general tendency in all countries to underestimate the magnitude of the effort required to programme and install computer systems. Because of the interaction of computer processing with virtually all other elements of the census, it is essential that the computer programmes be prepared and fully tested beforehand. It is not sufficient for the programmes merely to be ready by the time the main processing is scheduled to begin.

23. The more difficult part of the programming lies in the machine editing procedures, that is, in having the computer check for inconsistencies such as improbable code combinations. Translating the editing rules into steps sufficiently simple and precise for the computer to follow is usually more difficult than anticipated. It is of particular importance to check out these programmes well in advance using realistic data. Machine editing requires that non-responses and inconsistencies be detected and corrected by having the computer impute the corrected data in so far as possible. Early trial runs of the editing programmes will disclose other possible phases of the census operation which are the likely sources of editing problems. As a result, the need for corrective action in such areas as the enumeration form, enumerator instructions, coding, key-punching or supervision may be brought into prominence.

24. If thorough and early testing of the complete data processing phase is not conducted, it is likely that during the actual operations the computer will reject excessively large numbers of records with one defect or another. This, in turn, creates the problem of reinstating the records that have been rejected and corrected as a result of the editing operation. The effect is likely to be unexpectedly time consuming, requiring more computer time than originally planned for.

25. Pre-testing of the entire range of computer operation, that is, the edit and tabulation programmes, will also provide much better advance estimates of the computer hours needed. Estimates based entirely on theoretical factors have often proved to be greatly in error. This consideration is important where an installation is heavily loaded and precise scheduling is necessary.

26. A great deal of time is required for programming for a computer and the lack of experienced staff is a common limiting factor in new computer installations. It requires trained personnel, and their training may have to be as much as one year for adequate proficiency. Thus, it is especially important that sufficient time be allowed in the census plan to permit the training of programmer staff early enough for completion and testing of the edit and tabulation programmes prior to the census date so that full-scale processing can begin immediately after the enumeration.

VI. CONCLUSION

27. The increased power and versatility of the present generation of machines over their predecessors will give those statisticians processing the 1970 census substantial advantages as compared with 1960. This equipment will introduce a marked improvement in the census procedures of the countries of the region, as did the introduction of punched-card equipment years ago. For efficient use of the equipment, the design of the processing needs to be integrated with the other aspects of the over-all census design. Integration of the capabilities of computer processing into the census operations as a whole entails a balance in supervision, quality control, editing, and the capacity to rectify detected errors. Imbalances may lead to serious problems.

28. It is urged, therefore, that the computer programming be completed well in advance of actual processing and that very thorough testing be performed on data obtained through pilot-tests. Though the outcome of this possibly may be an unchanged computer system design, it may well lead to strengthening of supervision and quality controls in the key stages of the census.

29. With the general application of computers to the 1970 round of population censuses and with careful planning of all stages of processing, administrators and planners of the censuses in the ECAFE region may look forward with reasonable confidence to reducing the interval between collection of data and publication to at least half the time required for the 1960 censuses.

. . . . .